

A screened automated structural search with semiempirical methods

Yukihiro Ota^a, Sergi Ruiz-Barragan^{a,b}, Masahiko Machida^a, Motoyuki Shiga^a

^aCCSE, Japan Atomic Energy Agency, 178-4-4 Wakashiba, Kashiwa 277-0871, Japan

^bDepartment of Theoretical and Computational Molecular Science, Institute of Molecular Science, Okazaki 444-8585, Japan

Abstract

We developed an interface program between a program suite for an automated search of chemical reaction pathways, GRRM, and a program package of semiempirical methods, MOPAC. A two-step structural search is proposed as an application of this interface program. A screening test is first performed by semiempirical calculations. Subsequently, a reoptimization procedure is done by ab initio or density functional calculations. We apply this approach to ion adsorption on cellulose. The computational efficiency is also shown for a GRRM search. The interface program is suitable for the structural search of large molecular systems for which semiempirical methods are applicable.

1. Introduction

A systematic exploration of chemical reaction pathways is one of the challenging issues in modern computational chemistry. This task is common to many computational calculations, such as the design of catalytic activity [1] and the creation of crystal structures from chemical composition alone [2]. Typically, the relevant computational task is to optimize a multi-dimensional function on potential energy surfaces [3] to find all the important local minima and saddle points. The well-ordered manipulation of a large amount of data is required as well. Significant research efforts have devoted to these topics [4, 5, 6]. A program suite, GRRM [6, 7], is intended to achieve an automated search of reaction pathways. This program suite has been applied to various issues in chemistry and materials science, such as the reaction pathways of oxygen atom on silicon surfaces [8], exploring conical intersections near the Franck-Condon region in different molecules [9], a bolylation of organic halides with silyboranes [10], and the prediction of undiscovered carbon structures [11].

The GRRM program enables an automated search of reaction routes. However, an efficient search is mandatory for treating large systems, such as polymers, proteins and biological molecules. The search algorithm in GRRM requires the calculation of forces upon the atoms of a target molecule along the potential energy surface [6], calculated by ab initio molecular orbital theory or density functional theory (DFT) with external program packages, such as GAUSSIAN 09 [12] and GAMESS [13, 14]. Thus, depending on available com-

putational resource, it is desirable to reduce the computational effort of force calculations.

In this article, we describe an interface program between GRRM and MOPAC [15, 16] to implement an automated search of chemical reaction pathways of large molecular system with semiempirical methods. This program is compatible with GRRM 14 [7]. The output data of MOPAC is converted into a readable format for input to GRRM 14, and vice versa. Moreover, we propose a two-step procedure to find stable structures in large molecular systems. The first step is a screening test via GRRM with semiempirical methods, to find the candidates for stable structures. The next step is the separate reoptimization of the resultant candidates with more costly but precise methods either based on ab initio or density functional calculations.

The interface program and the two-step structural search are tested for ion adsorption on cellulose. The mechanism of ion adsorption on materials and organic products is important for various industrial and environmental issues, such as designing nanosensors for hydrate fissile ions in waste water [17], the applications of graphene oxides to lithium-ion batteries [18], and the transport mechanism of radioactive cesium ion in plants [19]. We search for the stable structures of cellulose with three different alkali cations, Li^+ , K^+ , and Cs^+ . We also show a way of quantifying the relative stability of cation binding to molecules, relevant to measurable data in experiments. The efficiency of the present approach is discussed from the viewpoint of computational performance. We show that in a screening test with GRRM 14 the present approach

significantly reduces the computational time compared to GAMESS. Although semiempirical methods implemented in GAUSSIAN have been used previously with GRRM [20], it is much more efficient to use MOPAC directly, as will be evident from the present paper. Thus, as long as the semiempirical method chosen is qualitatively correct, the method proposed herein is highly desirable for large molecular systems.

2. Methods and computational details

Our interface program connecting between GRRM and MOPAC has been completed for the combination of GRRM 14 and MOPAC 2012. GRRM 14 includes two kinds of main algorithms for searching chemical reaction pathways. One is the anharmonic downward distortion following method, leading to an exploration of isomerization and dissociation pathways. The other is the artificial force induced reaction (AFIR) method, leading to a search of associative pathways of two or more reactants by way of transition states. Both methods require forces on the potential energy surfaces of target systems. When calling the interface program, the forces are calculated by a semiempirical method. Then, the standard output data of MOPAC is transformed into that readable for GRRM. Thus, an automated search of chemical reaction pathways is performed by a semiempirical method.

We search for the stable structures of α -cellulose $(C_6H_{10}O_5)_n$ ($n = 1, 2$) binding a single cation, as an application of the present interface program. The structure of α -cellulose is built up from crystal structure data [21] at the B3LYP [22, 23]/6-31G** level of theory by Gaussian 03. The edges of a chain structure of cellulose are terminated by $-OH$ and $-H$ assuming the product of hydrolysis reaction. The structure with $n = 1$ corresponds to *D*-glucose, but we call it cellulose monomer for convenience throughout this article. Three kinds of alkali cations, Li^+ , K^+ , and Cs^+ , are studied. A screening test is first performed by the multicomponent AFIR (MC-AFIR) method [24, 25] with PM6 [26] in MOPAC 2012. The MC-AFIR method in GRRM 14 allows us to produce different associative pathways from the randomly-generated configurations of several reactants, with the aid of multiple artificial forces between intra- and inter-reactant components. We apply the artificial force between the cation and each oxygen atom of cellulose, with the upper bound of a collision energy parameter [6] set to be 100 kJ/mol. The stopping criterion in this search is that an identical reaction pathway is discovered 50 times in a row. Subsequently, the structures

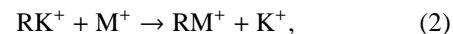
discovered in the above screening procedure are separately reoptimized at the B3LYP/LanL2DZ level of theory by GAUSSIAN 09. Each structure has an identification number in ascending order, according to increasing energy. The relative stability of structures is assessed by the difference of calculated energy, $E^{(s)} - E^{(s=1)}$ with the identification number of structures s and the calculated energy $E^{(s)}$.

After the search of stable structures, we study a solvent effect in the cation binding to cellulose molecules, within the polarizable continuum model (PCM) [27, 28]. The stable structures in aqueous solutions are calculated at the B3LYP/LanL2DZ level of theory by DFT calculations with the PCM of water. The above gaseous-phase results are utilized as initial structures. To study statistical properties in aqueous solutions at a temperature T , we use the canonical ensemble of stable structures. We may define the partition function as

$$Z_{\text{PCM}} = \sum' e^{-E_{\text{PCM}}^{(s)}/k_B T}, \quad (1)$$

with the Boltzmann constant k_B . The prime symbol on summation means that the summation index s runs over a set of distinct structures. We pick up distinct structures using inter-atomic distance; if all the inter-atomic distances of the s th stable structure are equal to those of the s' th structure, the two structures are considered to be identical. The threshold value of distance is set as 0.1 Å. The canonical ensemble leads to the probability of finding a certain structure. In this article, we focus on the probability of finding the most stable structure, $p_1 = Z_{\text{PCM}}^{-1} e^{-E_{\text{PCM}}^{(1)}/k_B T}$, at $T = 300$ K.

The two-step structural search allows us to obtain well-ordered data of cation-binding patterns, leading to an estimation of experimental data on ion adsorption, such as the amount of adsorption and the rate of ion exchange. To address this issue, let us focus on a reaction process



where R represents either cellulose monomers or dimers and M^+ does a cation. Calculating an energy difference between the reactants and the products would quantify the relative amount of cation adsorption on molecules. The energy data of discovered structures is suitable for estimating this difference. In a gaseous phase we can write this quantity as

$$\Delta E_{\text{gas}} = [E_{\text{DFT}}^{(s=1)}(RM^+) + E_{\text{DFT}}(K^+)] - [E_{\text{DFT}}^{(s=1)}(RK^+) + E_{\text{DFT}}(M^+)], \quad (3)$$

with the DFT-calculation energy of an isolated cation, $E_{\text{DFT}}(M^+)$.

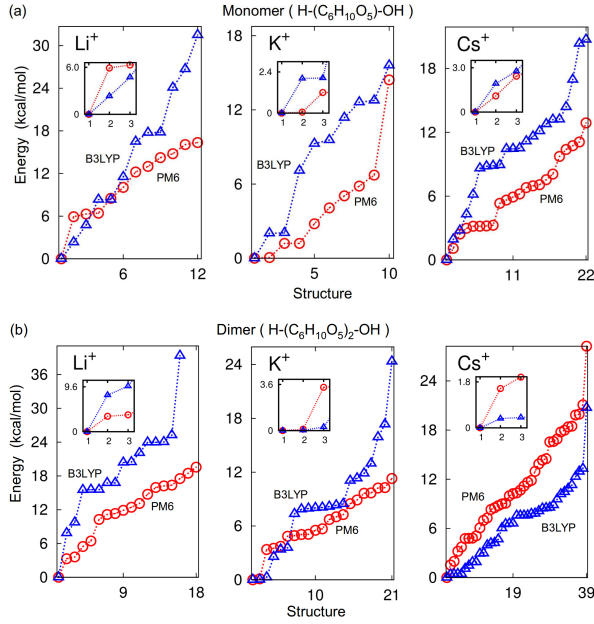


Figure 1: Energy differences between the stable structures of (a) cellulose monomers and (b) cellulose dimers with a cation, discovered by the multicomponent artificial force induced reaction (MC-AFIR) method [24, 25] with PM6 (red circles) and reoptimized by DFT calculations at the B3LYP/LanL2DZ level of theory (blue triangles). The base point of energy is the energy of Structure 1 (i.e. the lowest energy). Three kinds of cations, Li^+ , K^+ , and Cs^+ , are considered. The insets show the three lowest energies.

Since the experiments of ion adsorption on molecules are typically performed in water [19], a way of estimating the energy difference in aqueous solutions, denoted by ΔE_{aq} , is desirable for linking theoretical and experimental data. The two-step structural search with PCM leads to useful data on this issue. On the energy of RM^+ in Eq. (3), we evaluate the free energy according to partition function (1), $A_{\text{PCM}} = -k_B T \ln Z_{\text{PCM}}$, at $T = 300$ K rather than the energy of the most stable structure, to take the occurrence of different structures in aqueous solutions into account. The entropy, $S_{\text{PCM}} = -\partial_T A_{\text{PCM}}$, is also evaluated, to quantify the number of states in the statistical distribution of stable structures in aqueous solutions. As for an isolated cation, we replace $E_{\text{DFT}}(\text{M}^+)$ in Eq. (3) with the energy obtained by DFT calculations with PCM.

The efficiency of a GRRM search with MOPAC is examined by measuring the computational time for discovering the stable structures of cation binding to cellulose compared to GAMESS. Moreover, measuring a force-calculation time of DFT calculations, we estimate the computational time of a GRRM structural search with DFT.

3. Results and discussion

First, we study the relative stability of the structural data in the two-step search. Figure 1 shows the energy difference between the stable structures discovered by the MC-AFIR method with PM6 (red circles) and reoptimized by DFT calculations with B3LYP/LanL2DZ (blue triangles). The horizontal axes indicate the structure identification numbers, while the vertical axes show the energy difference between different structures. The insets show the data of the three lowest energies. We obtain cation binding structures with a wide range of energy. Let us focus on the reoptimization results (blue triangles). We find that in cellulose dimers multiple lowest-energy structures appear when the atomic number of cations increases (i.e. the radius of cations becomes large). As for K^+ [middle panel of Figure 1(b)], we have the two distinct structures within 0.6 kcal/mol. We count the number of distinct stable structures in the manner of checking identical inter-atomic distances with threshold distance 0.1 Å. Similarly, as for Cs^+ [right panel of Figure 1(b)], we have the three distinct lowest-energy structures. Otherwise, the lowest-energy structures are well separated from the higher-energy ones.

Next, we show the cation binding structures of cellulose molecules. Figure 2 shows the stable structures with cation-oxygen distances, mainly focusing on the most stable structures in DFT calculations with PCM. All the analyses of molecular structures and visualization were performed in VMD [29]. On the bottom of each panel we show the probability of finding the corresponding structure according to the canonical ensemble described by partition function (1).

In Figure 2(a), we show a typical sequence of structures from screening to reoptimization on cellulose dimers with Li^+ , as well as the most stable structure in DFT calculations with PCM (right panel). The identification numbers on the bottom of the left and middle panels correspond to those in the horizontal axes of Figure 1. The most stable structure in DFT calculations with PCM comes from the 11th stable structure discovered by the MC-AFIR method with PM6. Thus, the reoptimization by DFT calculations can alter the stability order of the structures predicted by semiempirical methods in screening. We stress, however, that the use of GRRM with MOPAC leads to different kinds of the initial guesses for subsequent high-level calculations in an unbiased and automated way, rather than the precise data on molecular systems.

We turn into the probability of finding the most stable structure in aqueous solutions in Figure 2. The aque-

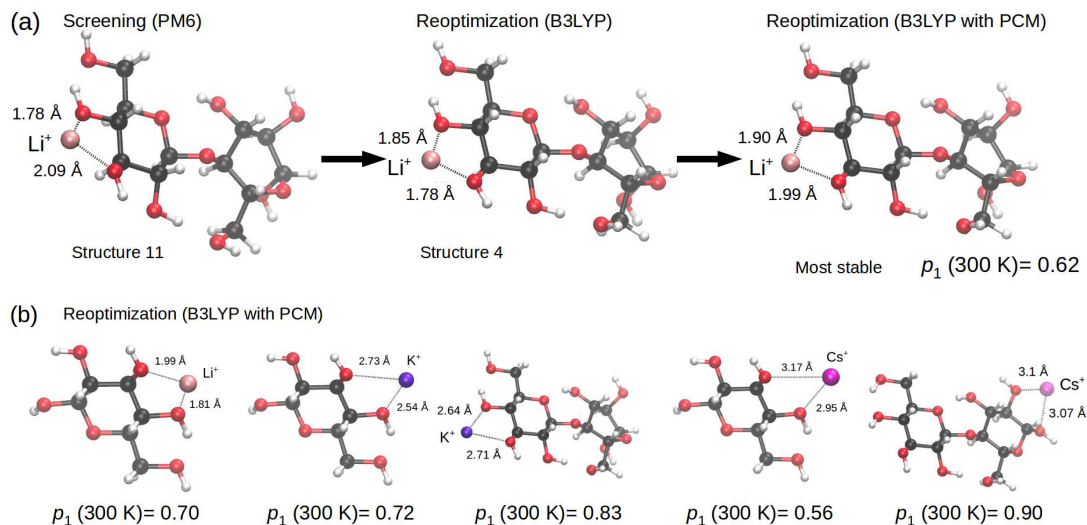


Figure 2: (a) Sequence of stable structures on cellulose dimers binding Li^+ . The most stable structure obtained by DFT calculations with PCM (right panel) comes from the 11th stable structure (left panel) in the screening performed by the MC-AFIR method with PM6, via the 4th stable one (middle panel) obtained by DFT calculations without PCM (i.e. gaseous phase). The identification numbers on the left and middle panels correspond to those in the horizontal axes of Figure 1. On the bottom of the right panel, the probability of finding the most stable structure according to partition function (1) p_1 at 300 K is shown. (b) Most stable structures and finding probability p_1 at 300 K obtained by DFT calculations with PCM, on cellulose monomers and dimers binding a cation. The result of a cellulose dimer with Li^+ is shown in the right panel of (a).

ous solutions described by PCM can change the relative stability between stable structures in the gaseous phase. We find drastic changes for Cs^+ . On cellulose monomers binding Cs^+ , the probability is less than 0.60. Thus, solvents lead to a broad distribution of stable structures, although in the gaseous phase a single lowest-energy structure is well separated from higher-energy ones [right panel of Figure 1(a)]. On the other hand, the cellulose dimers binding Cs^+ have a single lowest-energy structure with very high probability ($p_1 = 0.9$). This result is in contrast to the gaseous-phase ones in Figure 1, where there are multiple lowest-energy structures within 0.6 kcal/mol. The results for entropy [lower panel of Figure 3(b)] also indicate these effects.

Now, we study the amount of cation adsorption on cellulose molecules, according to reaction process (2). Figure 3 shows the energy differences, ΔE_{gas} and ΔE_{aq} , between the binding of different cations. On the lower panel of Figure 3(b), the entropy of stable structures is also shown. In both the gaseous-phase [Figure 3(a)] and aqueous-solution results [Figure 3(b)] cellulose molecules favor binding Li^+ . Moreover, the energy difference increases monotonically with the radius of the cation. However, the variation range of ΔE_{aq} significantly narrows, compared to that of the gaseous-phase results. Thus, the aqueous solutions described by PCM

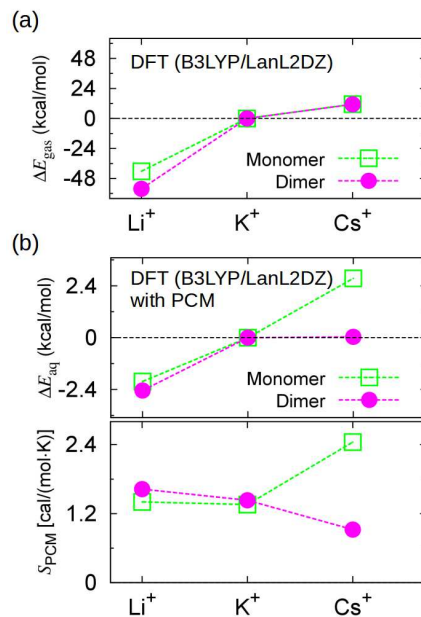


Figure 3: Energy difference associated with reaction process (2), (a) in the gaseous phase and (b) in aqueous solution. The energy differences in the gaseous phase contain the DFT-calculation energy of the most stable structure, whereas in aqueous solution they have the free energy relevant to the canonical ensemble of stable structures. The evaluation way is explained in the main text (Section 2). A negative value indicates that a cation is more strongly bound to a cellulose molecule than K^+ . Moreover, on the lower panel of (b), the entropy of stable structures according to partition function (1) is shown.

Table 1: Statistical information in a search of cation binding to cellulose monomers by GRRM14 with PM3, depending on the computational packages of force calculations. The search is stopped when the number of equilibrium structures reaches 10. The number of force calculations (n_{force}) and the total elapsed time (t_{elapsed}) are measured. Then, the mean calculation time of force is evaluated by $\bar{t}_{\text{force}} = t_{\text{elapsed}}/n_{\text{force}}$. Since in GAMESS the PM3 parameter set of Cs^+ is absent, the entries are empty. For reference, a force-calculation time of DFT calculations at the B3LYP/LanL2DZ level of theory is also shown in the last column of the second row. The structural data in the single-point DFT calculations is built up by adding a single cation to the optimized structure of a cellulose monomer at the B3LYP/LanL2DZ level of theory. The distance between a cation and the center of mass of a monomer is set as about 5 Å.

Packages of force calculations	Cations	n_{force}	t_{elapsed} (sec.)	\bar{t}_{force} (sec.)	$t_{\text{force(DFT)}}$ (sec.)
MOPAC	Li^+	7406	670	0.09	—
	K^+	7121	610	0.09	—
	Cs^+	5055	430	0.09	—
GAMESS	Li^+	5501	5261	0.96	135.2
	K^+	4473	3880	0.87	151.8
	Cs^+	—	—	—	140.9

lead to a reduction in the relative energy costs associated with cation binding to cellulose. This reduction comes purely from the change in electrostatic energy of molecules since the contributions from the entropy are quite small as seen in Figure 3(b). Thus, the two-step structural search is useful for studying the distribution of stable structures in molecules.

Now, we show the efficiency of our approach. The computational costs are evaluated on a desktop machine with the Intel[®] Xeon[®] E5645 processor, compared to the use of GAMESS. We employ the PM3 model [30, 31]. In the screening processes done by GRRM, the number of force calculations, n_{force} , and the total elapsed time of search, t_{elapsed} , are measured. Then, the mean time of force calculations, $\bar{t}_{\text{force}} = t_{\text{elapsed}}/n_{\text{force}}$, is estimated. The screening processes are stopped when the number of discovered equilibrium structures reaches 10. Table 1 shows the statistical information in a search of Li^+ binding to cellulose monomers, by the MC-AFIR method with PM3. Since in GAMESS the PM3 parameter set of Cs^+ is absent, the entries are empty. For reference, a single-point computational time of force calculations by DFT via GAMESS is also shown in the last column. The table indicates that in a search by GRRM a use of MOPAC is much more efficient than that of GAMESS. In addition, we find that a GRRM search with MOPAC via the interface program is about 10^3 times faster than the use of DFT force calculations if the number of force calculations are common to the two force-calculation methods. Thus, our interface program is useful for automatically producing various stable structures in large-scale molecules.

Finally, we discuss a range of the applications of the interface program. The validity of our approach depends on that of semiempirical methods. Therefore, the use of the interface program is not suitable for discover-

ing the stable structures of molecular systems with transition metals and searching chemical reaction pathways including the rearrangement of covalent bonds. The transition-state search should be avoided, as well. In contrast, a stable-structure search in organic products is a good target for our interface program. A development of semiempirical methods would extend the application range.

4. Conclusion

We constructed an interface program between GRRM and MOPAC, to implement an automated search of stable structures in large-scale molecules with semiempirical methods. We applied this program to studying cation binding to cellulose monomers and dimers. Our approach is a two-step way of discovering stable structures. After a search of stable structures by GRRM with PM6, the resultant structures were reoptimized at the B3LYP/LanL2DZ level of theory by DFT. We found the cation binding structures with a wide range of energy. We also demonstrated a way of estimating experimental data on ion adsorption using well-ordered data of different structures in aqueous solutions within the PCM of water. Moreover, the efficiency of a GRRM search with the interface program was shown, compared to the use of GAMESS to calculate forces with semiempirical methods. The use of GRRM with MOPAC leads to different kinds of the initial guesses for high-level calculations in an unbiased and automated way. The present interface program is applicable to various chemical-reaction-search issues in large-scale molecules, within the validity of semiempirical methods.

Acknowledgments

We would like to thank S. Maeda for his helpful comments. M.M. acknowledges fruitful discussion with T. Doi and her collaborators. This work is partially supported by Sector of Fukushima Research and Development in JAEA. We thank M. Yui and his colleagues for their support. The calculations were partially performed on JAEA BX900 supercomputer. We thank CCSE staff members for their assistance.

References

- [1] K. N. Houk, P. H.-Y. Cheong, *Nature* **455** (2008) 309.
- [2] S. M. Woodley, R. Catlow, *Nat. Mater.* **7** (2008) 937
- [3] F. Jensen, *Introduction to Computational Chemistry*, Second Edition, John & Wiley, West Sussex, 2007.
- [4] D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.
- [5] A. O. Lyakhov, A. R. Oganov, H. Stokes, Q. Zhu, *Comput. Phys. Commun.* **184** (2013) 1172.
- [6] S. Maeda, K. Ohno, K. Morokuma, *Phys. Chem. Chem. Phys.* **15** (2013) 3683.
- [7] S. Maeda, Y. Harabuchi, Y. Osada, T. Taketsugu, K. Morokuma, K. Ohno, see <http://grrm.chem.tohoku.ac.jp/GRRM> (accessed on 11 Apr., 2014).
- [8] S. Ohno, K. Shudo, M. Tanaka, S. Maeda, K. Ohno, *J. Phys. Chem. C* **114** (2010) 15671.
- [9] S. Maeda, Y. Harabuchi, T. Taketsugu, K. Morokuma, *J. Phys. Chem. A* **118** (2014) 12050.
- [10] R. Uematsu, E. Yamamoto, S. Maeda, H. Ito, T. Taketsugu, *J. Am. Chem. Soc.* **137** (2015) 4090.
- [11] K. Ohno, H. Satoh, T. Iwamoto, H. Tokoyama, H. Yamakado, *Chem. Phys. Lett.* **639** (2015) 178.
- [12] Gaussian 09, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian, Inc., Wallingford CT, 2010.
- [13] General Atomic and Molecular Electronic Structure System, M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.* **14** (1993) 1347.
- [14] M. S. Gordon, M.W.Schmidt, *Advances in electronic structure theory: GAMESS a decade later*, pp. 1167-1189, in *Theory and Applications of Computational Chemistry: the first forty years*, C. E. Dykstra, G. Frenking, K. S. Kim, G. E. Scuseria (Ed.), Elsevier, Amsterdam, 2005.
- [15] MOPAC 2012, J. J. P. Stewart, *Stewart Computational Chemistry*, Version 15.180L. <http://www.openmopac.net/index.html>
- [16] J. D. C. Mala, G. A. U. Carvalho, C. P. Manguiera Jr., S. R. Santana, L. A. F. Cabral, G. B. Rocha, *J. Chem. Theory Comput.* **8** (2012) 3072.
- [17] N. Kumar, J. M. Seminario, *J. Phys. Chem. C* **117** (2013) 24033.
- [18] C. Uthaisar, D. J. Hicks, V. Barone, *Surface Science* **619** (2014) 105.
- [19] E. Adams, V. Chaban, H. Khandelia, R. Shin, *Sci. Rep.* **5** (2015) 8842.
- [20] T. Takayanagi, K. Takahashi, A. Kakizaki, M. Shiga, M. Tachikawa, *Chem. Phys.* **358** (2009) 196.
- [21] Y. Nishiyama, P. Langan, H. Chanzy, *J. Am. Chem. Soc.* **124** (2002) 9074.
- [22] A. D. Becke, *J. Chem. Phys.* **98** (1993) 5648.
- [23] P.J. Stephens, F.J. Devlin, C.F. Chabalowski, M.J. Frisch, *J. Phys. Chem.* **98** (1994) 11623.
- [24] S. Maeda, K. Morokuma, *J. Chem. Phys.* **132** (2010) 241102.
- [25] S. Maeda, K. Morokuma, *J. Chem. Theory Comput.* **7** (2011) 2335.
- [26] J. J. P. Stewart, *J. Mol. Model.* **13** (2007) 1173.
- [27] J. Tomasi, M. Persico, *Chem. Rev.* **94** (1994) 2027.
- [28] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **105** (2005) 2999.
- [29] W. Humphrey, A. Dalke, K. Schulten, *J. Molec. Graphics* **14** (1996) 33.
- [30] J. J. Stewart, *J. Comp. Chem.* **10** (1989) 209.
- [31] J. J. Stewart, *J. Comp. Chem.* **10** (1989) 221.